# Classification automatique de mesures de similarité

Fabien Duchateau, Franck Favetta (prénom.nom@liris.cnrs.fr)

**Contexte.** L'appariement d'entités consiste à identifier des instances équivalentes dans des jeux de données [1]. Par exemple, dans les tableaux ci-dessous, les instances *123* (à gauche) et *P1* (à droite) correspondent au même objet du même réel (i.e., le roman *les deux tours*). Pour détecter ces entités correspondantes, on utilise des mesures de similarité qui calculent un score entre les propriétés des entités à comparer. Par exemple, le coefficient de Dice calcule un score de 0,51 entre les chaînes *Holy Grail* et *Holy Grail - Dark Passenger*. Il existe de nombreuses mesures de similarité [2, 3, 4], qui sont généralement catégorisées selon leur fonctionnement. Dans ce projet, nous voulons classer les mesures selon leurs résultats.

| id | nom | type |
|---|---|---|
| 123 | Les deux tours | livre |
| 345 | Holy grail | DVD |
| 567 | Les deux tours | DVD |

| id | titre | catégorie |
|---|---|---|
| P1 | Le seigneur des anneaux - Les 2 tours | roman |
| P2 | Holy Grail - Dark Passenger | musique |

**Objectifs.** Ce projet cherche à atteindre les objectifs suivants :

- Rédiger un court état de l'art sur les mesures de similarité et leur classification;
- Améliorer un script Python existant qui calcule des scores de similarité, pour un ensemble de mesures (e.g., Talisman, Jellyfish) et un jeu de données;
- Développer un script Python pour classifier les mesures de similarité en utilisant des algorithmes de clustering (e.g., scikit-learn);
- Étudier les résultats obtenus (comparaison avec la classification traditionnelle, robustesse de la classification, configuration des paramètres comme le nombre de catégories, etc.).

**Bibliographie.**

[1] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, 2015. Lien TKDE14.

[2] Seung seok Choi and Sung hyuk Cha. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pages 43–48, 2010. Lien JSCI2010.

[3] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava. Record linkage: similarity measures and algorithms. In *SIGMOD*, pages 802–803. ACM, 2006. Lien SM2006.

[4] Najlah Gali, Radu Mariescu-Istodor, and Pasi Fränti. Similarity measures for title matching. In *2016 23rd International Conference on Pattern Recognition*, pages 1548–1553. IEEE, 2016. Lien ICPR2016.